Supporting Information

Systematic Protein Prioritization for Targeted Proteomics Studies through Literature Mining

Kun-Hsing Yu[1,2], Tsung-Lu Michael Lee[3], Chi-Shiang Wang[4], Yu-Ju Chen[5], Christopher Ré[6], Samuel C. Kou[2], Jung-Hsien Chiang[4,], Isaac S. Kohane[1,*], Michael Snyder[7,*]*

[1] Department of Biomedical Informatics, Harvard Medical School

[2] Department of Statistics, Harvard University

[3] Department of Information Engineering, Kun Shan University, Taiwan

[4] Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan

[5] Institute of Chemistry, Academia Sinica, Taiwan

[6] Department of Computer Science, Stanford University

[7] Department of Genetics, Stanford University

Table of Contents:

Table S-1. Query terms for the 22 B/D-HPP targeted areas.

B/D HPP Targeted Areas	Query Terms
Brain	brain
Cancer	cancer
Cardiovascular	cardiovascular
Diabetes	diabetes
EyeOME	eye OR ocular
Food and nutrition	food OR nutrition OR nutrients
Glycoproteomics	glycoproteins
Immune-peptidome	immune OR immune system
Infectious diseases	infectious OR infection
Kidney and urine	kidney OR urine
Liver	liver OR hepatic
Mitochondria	mitochondria
Model organisms	rat OR mouse
Musculoskeletal	muscle OR bone OR musculoskeletal
Pathology	pathology
PediOme	pediatric OR newborn OR infant OR toddler OR child OR adolescent
Plasma	plasma OR serum
Protein aggregation	protein aggregation

Rheumatic disorders	rheumatic
Stem cells	stem cells
Toxicoproteomics	toxicology OR toxic OR toxin
Extreme conditions	hot OR cold OR alkaline condition OR acidic condition OR hypersaline OR radiation

Table S-2. Most frequent genetic mutations co-published with the common cancer types.

	Cancer	Lung	Prostate	Colorectal	Stomach	Liver	Bladder	Esophageal		Kidney		Breast	Cervical	Uterine	Ovarian	Thyroid	Brain
Rank	Overall	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Lymphoma	Cancer	Leukemia	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer
1	EGFR	EGFR	AR	KRAS	TP53	TP53	TP53	TP53	ALK	WT1	ABL1	BRCA1	TP53	TP53	BRCA1	RET	EGFR
2	TP53	TP53	KLK3	EGFR	CDH1	KRAS	FGFR3	EGFR	EGFR	VHL	FLT3	ERBB2	EGFR	MSH6	TP53	BRAF	IDH1
3	BRCA1	KRAS	BRCA1	MLH1	ERBB2	EGFR	CDKN2A	CDKN2A	ETV6	TP53	KMT2A	TP53	CDKN2A	BRCA1	ERBB2	PTEN	TP53
4	KRAS	ALK	TP53	MSH6	KIT	CTNNB1	ERBB2	ERBB2	ABL1	TFE3	TP53	EGFR	MYC	KRAS	PARP1	EGFR	CDKN2A
5	CDKN2A	CDKN2A	PTEN	TP53	EGFR	AFP	EGFR	KRAS	MYC	FH	AGXT	ESR1	ERBB2	FH	KRAS	HRAS	BRAF
6	ERBB2	ERBB2	EGFR	CTNNB1	MLH1	CDKN2A	HRAS	CCND1	CDKN2A	FLCN	ETV6	CHEK2	RET	PTEN	BRCA2	TP53	MEN1
7	ABL1	BRAF	BRCA2	APC	KRAS	ERBB2	ZNF77	FHIT	TP53	EGFR	JAK2	BRCA2	FHIT	MLH1	EGFR	TSHR	PTEN
8	BRAF	MYC	CHEK2	MUTYH	CDKN2A	HFE	TERT	MLH1	BCL2	MTOR	RUNX1	PTEN	KRAS	ERBB2	MLH1	GNAS	CTNNB1
9	RET	FHIT	ERBB2	MSH2	CTNNB1	KIT	KRAS	KIT	KMT2A	CTNNB1	CEBPA	PIK3CA	PIK3CA	CTNNB1	CHEK2	MEN1	CDK4
10	MLH1	CTNNB1	HOXB13	BRAF	MSH6	BRAF	RB1	PIK3CA	MME	CDKN2A	FANCB	CDKN2A	BRAF	EGFR	MSH6	KRAS	VHL
11	CTNNB1	CYP1A1	ERG	CDKN2A	PDGFRA	MYC	MSH6	CDKN1A	FLT3	SMARCB1	PML	ATM	RB1	CDKN2A	CTNNB1	NRAS	SMARCB1
12	KIT	AKT1	CDKN2A	DCC	RBBP4	CDKN1A	NAT2	PTEN	JAK2	KRAS	CDKN2A	PARP1	PDXP	ESR1	PTEN	PTCH1	MSH6
13	FLT3	HRAS	AKT1	RBBP4	BRCA1	HCCS	MLH1	FGF3	IKZF1	TSC1	NPM1	AKT1	HRAS	PIK3CA	AKT1	CTNNB1	MGMT
14	JAK2	FGFR1	RNASEL	BRCA1	CDKN1A	DLC1	CDKN1A	CTNNB1	CCND1	PTEN	MYC	CDH1	BRCA1	MYC	CDKN2A	CDKN2A	NF1
15	MYC	KIT	CTNNB1	SMAD4	MALT1	AKT1	MSH2	CDK4	DDIT3	KRT7	CSF3	PGR	STK11	MED12	DICER1	ERBB2	GH1
16	NF1	CEACAM3	PARP1	ERBB2	SMAD4	HNF1A	GSTM1	APC	MALT1	EPAS1	ASXL1	KRAS	PTEN	AKT1	ESR1	PPARG	ERBB2
17	MSH6	FLCN	CDKN1A	CDKN1A	MSH2	PTEN	MYC	POLB	AGXT	IGF2	KIT	CDKN1A	MDM2	MSH2	PIK3CA	ZHX2	GNAS
18	RB1	DICER1	KLF6	PIK3CA	PTEN	PIK3CA	FGFR1	DCC	NOTCH1	PAX6	CSF2	CTNNB1	FGFR3	RB1	COL11A2	AKT1	MYC
19	PTEN	GSTM1	MYC	CEACAM3	GAST	CEACAM3	AKT1	FGF4	BCL6	PRCC	NRAS	RAD51	CDKN1A	FHIT	RAD51C	NF1	NF2
20	IDH1	RB1	KRAS	PTEN	DES	MSH6	CTNNB1	BRCA2	MYD88	HIF1A	DNMT3A	AR	AKT1	NLRP7	CDKN1A	TERT	AIP

Table S-3. Comparison of proteins associated with cancer, the cardiovascular system, diabetes, and the liver in human, rat, and mouse. Colored: common proteins with high ranks across species.

Rank	Cancer			Cardiovascular			Diabetes			Liver		
	Human	**Rat**	**Mouse**	**Human**	**Rat**	**Mouse**	**Human**	**Rat**	**Mouse**	**Human**	**Rat**	**Mouse**
1	**EGFR**	**Kras**	**Kras**	CRP	**Agt**	Apoe	**INS**	**Gcg**	**Adipoq**	**AFP**	Ccl4	Pcsk9
2	**ERBB2**	**Myc**	**Erbb2**	**ACE**	**Gja1**	**Agt**	**DPP4**	**Iapp**	**Gcg**	GPT	Ggt1	Fah
3	KLK3	**Egfr**	Braf	**CDH5**	**Akt1**	**Nos3**	**GCG**	**Akt1**	**Lep**	SLC17A5	Got2	Adipoq
4	**CDKN2A**	Bcl2	**Apc**	**EDN1**	**Edn1**	Gata4	CRP	**Gck**	Neurog3	SLCO1B1	**Gck**	Abca1
5	**CTNNB1**	Tsc2	**Egfr**	INS	Rhoa	Angpt1	**SLC5A2**	**Slc2a4**	**Insr**	IFNL3	Cat	Nr1h4
6	**AKT1**	**Akt1**	**Brca1**	COG2	Nox1	Ldlr	**IAPP**	Cat	**Akt1**	CYP1A2	Pdlim3	**Cyp2e1**
7	CD24	Gstp1	**Cdkn2a**	**AGT**	**Eln**	**Akt1**	**GLP1R**	Akr1b1	**Dpp4**	CYP3A4	Gsr	Nr1i3
8	**PARP1**	Cd44	Pten	NPPB	Bcl2	Atp2a2	**ADIPOQ**	**Adipoq**	**Ins2**	**NR1H3**	Scd1	**Afp**
9	ESR1	Gja1	**Parp1**	PCSK9	**Nos3**	Cybb	**GCK**	**Ins1**	Ager	PNPLA3	**G6pc**	Ahr
10	CD274	**Apc**	Atm	ICAM1	Mapk8	**Gja1**	ACE	**Insr**	**Gck**	INS	Hgf	**Ppara**
11	**BRCA1**	Ggt1	**Ctla4**	CD59	Nppa	Adipoq	PPARGC1A	Il1b	Lepr	CYP2B6	**Cyp1a1**	**Nr1h3**
12	**CDKN1A**	Afp	**Akt1**	DPP4	**Ace**	**Ryr2**	SLC30A8	Agt	**Glp1r**	**FABP1**	Timp1	**Gck**

13	ABL1	Bckdha	**Ctnnb1**	APOA1	Hmox1	**Kdr**	GAD2	Igf1	Ppara	HCCS	**Cyp2e1**	**G6pc**
14	**KRAS**	Il2	Birc5	FGB	Cat	**Cdh5**	CD59	**Pparg**	Irs1	ADAM17	**Ppara**	**Fabp1**
15	MTOR	**Cdkn1a**	Ccnd1	CETP	Slc8a1	Cldn5	TCF7L2	**Lep**	Apoe	YAP1	Mlxipl	Hnf4a
16	IDH1	Cyp1a1	Nf1	APOB	**Ryr2**	Ager	NAMPT	**Ins2**	**Slc2a4**	CTNNB1	Nr3c1	Arntl
17	YAP1	Il1b	**Myc**	**NOS3**	Nos2	**Eln**	FGF21	Mapk8	**Pparg**	CEACAM3	Cyp3a62	Srebf1
18	TENM1	Hras	Cdh1	DBP	Il1b	Abca1	**PPARG**	Slc2a2	Gcgr	**HFE**	**Afp**	**Hfe**
19	CEACAM3	Igf1	Runx1	NKX2-5	**Kdr**	**Edn1**	**LEP**	Bcl2	Pdx1	**CYP1A1**	**Cyp7a1**	Acta2
20	**CTLA4**	Ins1	Ar	CST3	Vegfa	Flt1	KCNJ11	**Slc5a2**	Cd4	SAMSN1	Insr	**Cyp7a1**

Table S-4. Comparison of the precision, recall, and F1 score of PURPOSE, a variant of PURPOSE without integrating citation counts, GLAD4U, and PubPular. The best performance in each category is highlighted in bold.

	Cardiovascular			Kidney			Liver			Lung		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
PURPOSE	**0.560**	**0.265**	**0.360**	**0.384**	**0.374**	**0.379**	**0.512**	**0.170**	**0.255**	**0.400**	**0.313**	**0.351**
PURPOSE without Integrating Citation Counts	0.530	0.251	0.341	0.354	0.344	0.349	0.498	0.165	0.248	0.370	0.290	0.325
GLAD4U	0.502	0.237	0.322	0.232	0.226	0.229	0.368	0.122	0.183	0.324	0.213	0.257
PubPular	0.352	0.167	0.227	0.166	0.161	0.164	0.256	0.085	0.128	0.168	0.132	0.148